Latent Alignment and Variational Attention

Yuntian Deng[*]   **Yoon Kim**[*]   **Justin Chiu**   **Demi Guo**   **Alexander M. Rush**

{dengyuntian@seas,yoonkim@seas,justinchiu@g,dguo@college,srush@seas}.harvard.edu

School of Engineering and Applied Sciences
Harvard University
Cambridge, MA, USA

Abstract

Neural attention has become central to many state-of-the-art models in natural language processing and related domains. Attention networks are an easy-to-train and effective method for softly simulating alignment; however, the approach does not marginalize over latent alignments in a probabilistic sense. This property makes it difficult to compare attention to other alignment approaches, to compose it with probabilistic models, and to perform posterior inference conditioned on observed data. A related latent approach, hard attention, fixes these issues, but is generally harder to train and less accurate. This work considers *variational attention* networks, alternatives to soft and hard attention for learning latent variable alignment models, with tighter approximation bounds based on amortized variational inference. We further propose methods for reducing the variance of gradients to make these approaches computationally feasible. Experiments show that for machine translation and visual question answering, inefficient exact latent variable models outperform standard neural attention, but these gains go away when using hard attention based training. On the other hand, variational attention retains most of the performance gain but with training speed comparable to neural attention.

1 Introduction

Attention networks [6] have quickly become the foundation for state-of-the-art models in natural language understanding, question answering, speech recognition, image captioning, and more [13, 70, 14, 12, 55, 69, 61, 54]. Alongside components such as residual blocks and long-short term memory networks, soft attention provides a rich neural network building block for controlling gradient flow and encoding inductive biases. However, more so than these other components, which are often treated as black-boxes, researchers use intermediate attention decisions directly as a tool for model interpretability [37, 1] or as a factor in final predictions [23, 58]. From this perspective, attention plays the role of a latent alignment variable [8, 32]. An alternative approach, hard attention [69], makes this connection explicit by introducing a latent variable for alignment and then optimizing a bound on the log marginal likelihood using policy gradients. This approach generally performs worse (aside from a few exceptions such as [69]) and is used less frequently than its soft counterpart.

Still the latent alignment approach remains appealing for several reasons: (a) latent variables facilitate reasoning about dependencies in a probabilistically principled way, e.g. allowing composition with other models, (b) posterior inference provides a better basis for model analysis and partial predictions than strictly feed-forward models, which have been shown to underperform on alignment in machine translation [33], and finally (c) directly maximizing marginal likelihood may lead to better results.

---

[*]Equal contribution.

The aim of this work is to quantify the issues with attention and propose alternatives based on recent developments in variational inference. While the connection between variational inference and hard attention has been noted in the literature [4, 35], the space of possible bounds and optimization methods has not been fully explored and is growing quickly. These tools allow us to better quantify whether the general underperformance of hard attention models is due to modeling issues (i.e. soft attention imbues a better inductive bias) or optimization issues.

Our main contribution is a *variational attention* approach that can effectively fit latent variable alignments while remaining tractable to train. We consider two variants of variational attention: *categorical* and *relaxed*. The categorical method is fit with amortized variational inference using a learned inference network to improve the training bound and policy gradient variance reduction baseline using soft attention. With an appropriate inference network (which conditions on the entire source/target), it can be used at training time as a drop-in replacement for hard attention. The relaxed version assumes that the alignment is sampled from a Dirichlet distribution and hence allows attention over multiple source elements.

Figure 1: Sketch of variational attention applied to neural machine translation. Two alignment distribution are shown, the blue prior p computed using known information, and the red variational posterior q taking into account future observations. Our aim is to use q to improve estimates of p and to support improved inference of z.

Experiments describe how to implement this approach for two major attention-based models: neural machine translation and visual question answering (Figure 1 gives an overview of our approach for machine translation). We first show that maximizing exact marginal likelihood can increase performance over soft attention. We further show that with variational (categorical) attention, alignment variables significantly surpass both soft and hard attention results without requiring much more difficult training. We further explore the impact of posterior inference on alignment decisions, and how latent variable models might be employed. Our code is available at https://github.com/harvardnlp/var-attn/.

Related Work Latent alignment has long been a core problem in NLP, starting with the seminal IBM models [9], HMM-based alignment models [65], and a fast log-linear reparameterization of the IBM 2 model [18]. Neural soft attention models were originally introduced as an alternative approach for neural machine translation [6], and have subsequently been successful on a wide range of tasks (see [13] for a review of applications). Recent work has combined neural attention with traditional alignment [16, 62] and induced structure/sparsity [42, 29, 38, 74, 46, 47], which can be combined with the variational approaches outlined in this paper.

In contrast to soft attention models, hard attention [69, 3] approaches use a single sample at training time instead of a distribution. These models have proven much more difficult to train, and existing works typically treat hard attention as a black-box reinforcement learning problem with log-likelihood as the reward [69, 3, 45, 24, 17]. Two notable exceptions are [4, 35]: both utilize amortized variational inference to learn a sampling distribution which is used obtain importance-sampled estimates of the log marginal likelihood [10]. Our method uses uses different estimators and targets the single sample approach for efficiency, allowing the method to be employed for NMT and VQA applications.

There has also been significant work in using variational autoencoders for language and translation application. Of particular interest are those that augment an RNN with latent variables (typically Gaussian) at each time step [15, 20, 57, 21, 34] and those that incorporate latent variables into sequence-to-sequence models [73, 7, 60]. Our work differs by modeling an explicit model component (alignment) as a latent variable instead of auxiliary latent variables (e.g. topics). One example [7] also use the term "variational attention" to refer to a different component the output from attention (commonly called the context vector) as a latent variable—in contrast we model the explicit attention alignment.
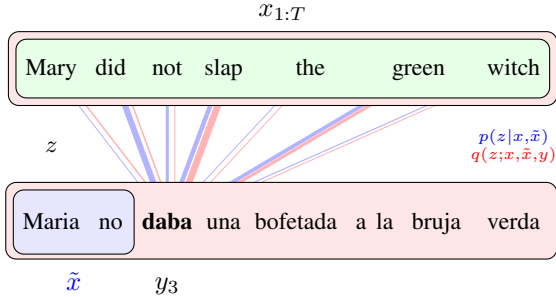
2 Background: Latent Alignment and Neural Attention

We begin by introducing notation for latent alignment, and then show how it relates to neural attention. For clarity, we are careful to use *alignment* to refer to this probabilistic model (Section 2.1), and *soft* and *hard* attention to refer to two particular inference approaches used in the literature to estimate alignment models (Section 2.2).

2.1 Latent Alignment

Figure 2(a) shows a latent alignment model. Let x be an observed set with associated members $\{x_1, \ldots, x_i, \ldots, x_T\}$. Assume these are vector-valued (i.e. $x_i \in \mathbb{R}^d$) and can be stacked to form a matrix $X \in \mathbb{R}^{d \times T}$. Let the observed $\tilde{x}$ be an arbitrary "query". These generate a discrete output variable $y \in \mathcal{Y}$. This process is mediated through a latent alignment variable z, which indicates which member (or mixture of members) of x generates y. The generative process we consider is:

$$z \sim \mathcal{D}(a(x, \tilde{x}; \theta)) \quad y \sim f(x, z; \theta)$$

where a produces the parameters for an alignment distribution $\mathcal{D}$. The function f gives a distribution over the output, e.g. an exponential family. To fit this model to data, we set the model parameters θ by maximizing the log marginal likelihood of training examples $(x, \tilde{x}, \hat{y})$:[2]

$$\max_{\theta} \ \log p(y = \hat{y} \,|\, x, \ \tilde{x}) \quad = \quad \max_{\theta} \ \log \mathbb{E}_z[f(x, z; \theta)_{\hat{y}}]$$

Directly maximizing this log marginal likelihood in the presence of the latent variable z is often difficult due to the expectation (though tractable in certain cases).

For this to represent an alignment, we restrict the variable z to be in the simplex Δ^{T-1} over source indices $\{1, \ldots, T\}$. We consider two distributions for this variable: first, let $\mathcal{D}$ be a *categorical* where z is a one-hot vector with $z_i = 1$ if x_i is selected. For example, $f(x, z)$ could use z to pick from x and apply a softmax layer to predict y, i.e. $f(x, z) = \mathrm{softmax}(\mathbf{W} X z)$ and $\mathbf{W} \in \mathbb{R}^{|\mathcal{Y}| \times d}$,
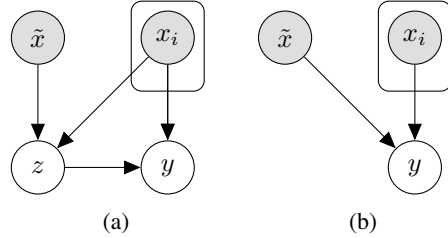

Figure 2: Models over observed set x, query $\tilde{x}$, and alignment z. (a) Latent alignment model, (b) Soft attention with z absorbed into prediction network.

$$\log p(y = \hat{y} \,|\, x, \ \tilde{x}) = \log \sum_{i=1}^{T} p(z_i = 1 \,|\, x, \tilde{x}) p(y = \hat{y} \,|\, x, z_i = 1) = \log \mathbb{E}_z[\mathrm{softmax}(\mathbf{W} X z)_{\hat{y}}]$$

This computation requires a factor of $O(T)$ additional runtime, and introduces a major computational factor into already expensive deep learning models.[3]

Second we consider a *relaxed* alignment where z is a mixture taken from the interior of the simplex by letting $\mathcal{D}$ be a Dirichlet. This objective looks similar to the categorical case, i.e. $\log p(y = \hat{y} \,|\, x, \ \tilde{x}) = \log \mathbb{E}_z[\mathrm{softmax}(\mathbf{W} X z)_{\hat{y}}]$, but the resulting expectation is intractable to compute exactly.

2.2 Attention Models: Soft and Hard

When training deep learning models with gradient methods, it can be difficult to use latent alignment directly. As such, two alignment-like approaches are popular: *soft attention* replaces the probabilistic model with a deterministic soft function and *hard attention* trains a latent alignment model by maximizing a lower bound on the log marginal likelihood (obtained from Jensen's inequality) with policy gradient-style training. We briefly describe how these methods fit into this notation.

---

[2]When clear from context, the random variable is dropped from $\mathbb{E}[\cdot]$. We also interchangeably use $p(\hat{y} \,|\, x, \tilde{x})$ and $f(x, z; \theta)_{\hat{y}}$ to denote $p(y = \hat{y} \,|\, x, \tilde{x})$.

[3]Although not our main focus, explicit marginalization is sometimes tractable with efficient matrix operations on modern hardware, and we compare the variational approach to explicit enumeration in the experiments. In some cases it is also possible to efficiently perform exact marginalization with dynamic programming if one imposes additional constraints (e.g. monotonicity) on the alignment distribution [72, 71, 50].

Soft Attention   Soft attention networks use an altered model shown in Figure 2b. Instead of using a latent variable, they employ a deterministic network to compute an expectation over the alignment variable. We can write this model using the same functions f and a from above,

$$\log p_{\text{soft}}(y \mid x, \ \tilde{x}) = \log f(x, \mathbb{E}_z[z]; \theta) = \log \text{softmax}(\mathbf{W} X \mathbb{E}_z[z])$$

A major benefit of soft attention is efficiency. Instead of paying a multiplicative penalty of $O(T)$ or requiring integration, the soft attention model can compute the expectation before f. While formally a different model, soft attention has been described as an approximation of alignment [69]. Since $\mathbb{E}[z] \in \Delta^{T-1}$, soft attention uses a convex combination of the input representations $X\mathbb{E}[z]$ (the *context vector*) to obtain a distribution over the output. While also a "relaxed" decision, this expression differs from both the latent alignment models above. Depending on f, the gap between $\mathbb{E}[f(x, z)]$ and $f(x, \mathbb{E}[z])$ may be large.

However there are some important special cases. In the case where $p(z \mid x, \tilde{x})$ is deterministic, we have $\mathbb{E}[f(x, z)] = f(x, \mathbb{E}[z])$, and $p(y \mid x, \tilde{x}) = p_{\text{soft}}(y \mid x, \tilde{x})$. In general we can bound the absolute difference based on the maximum curvature of f, as shown by the following proposition.

Proposition 1. *Define $g_{x,\hat{y}} : \Delta^{T-1} \mapsto [0, 1]$ to be the function given by $g_{x,\hat{y}}(z) = f(x, z)_{\hat{y}}$ (i.e. $g_{x,\hat{y}}(z) = p(y = \hat{y} \mid x, \tilde{x}, z))$ for a twice differentiable function f. Let $H_{g_{x,\hat{y}}}(z)$ be the Hessian of $g_{x,\hat{y}}(z)$ evaluated at z, and further suppose $\|H_{g_{x,\hat{y}}}(z)\|_2 \leq c$ for all $z \in \Delta^{T-1}, \hat{y} \in \mathcal{Y}\}$, and x, where $\| \cdot \|_2$ is the spectral norm. Then for all $\hat{y} \in \mathcal{Y}$,*

$$| p(y = \hat{y} \mid x, \tilde{x}) - p_{\text{soft}}(y = \hat{y} \mid x, \tilde{x}) | \leq c$$

The proof is given in Appendix A.[4] Empirically the soft approximation strategy works remarkably well in many cases, and often moves towards a sharper distribution with training. Alignment distributions learned this way often correlate with human intuition (e.g. word alignment in machine translation), although less well than more targeted models [33].[5]

Hard Attention   Hard attention is an approximate inference approach for latent alignment (Figure 2a) [69, 4, 45, 24]. Hard attention takes a single hard sample of z (as opposed to a soft mixture) and then backpropagates through the model. The approach is derived by two choices: First apply Jensen's inequality to get a lower bound on the log marginal likelihood, $\log \mathbb{E}_z[p(y \mid x, z)] \geq \mathbb{E}_z[\log p(y \mid x, z)]$, then maximize this lower-bound with policy gradients/REINFORCE [66] to obtain unbiased gradient estimates,

$$\nabla_\theta \mathbb{E}_z[\log f(x, z))] = \mathbb{E}_z[\nabla_\theta \log f(x, z) + (\log f(x, z) - B)\nabla_\theta \log p(z \mid x, \tilde{x})],$$

where B is a baseline (not depending on z) that can be used to reduce the variance of this estimator. To implement this approach efficiently, hard attention uses Monte Carlo sampling to estimate the expectation in the gradient computation. For efficiency, a single sample from $p(z \mid x, \tilde{x})$ is used, in conjunction with other tricks to reduce the variance of the gradient estimator (discussed more below) [69, 43, 44].

3   Variational Attention for Latent Alignment Models

Amortized variational inference (AVI, closely related to variational auto-encoders) [31, 53, 43] is a class of methods for efficient approximate latent variable inference, using learned inference networks. In this section we explore this technique for deep latent alignment models, and propose methods for *variational attention* that combine the benefits of soft and hard attention.

First note that the key approximation step in hard attention is to optimize a lower bound derived from Jensen's inequality. This gap could be quite large, contributing to poor performance.[6] Variational

---

[4]It is also possible to study the gap in finer detail by considering distributions over the inputs of f that have high probability under approximately linear regions of f, leading to the notion of *approximately expectation-linear* functions, which was originally proposed and studied in the context of dropout [40].

[5]Another way of viewing soft attention is as simply a non-probabilistic learned function. While it is possible that such models encode better inductive biases, our experiments show that when properly optimized, latent alignment attention with explicit latent variables do outperform soft attention.

[6]Prior works on hard attention have generally approached the problem as a black-box reinforcement learning problem where the rewards are given by $\log f(x, z)$. Ba et al. (2015) [4] and Lawson et al. (2017) [35] are the notable exceptions, and both works utilize the framework from [44] which obtains multiple samples from a learned sampling distribution to optimize the IWAE bound [10] or a reweighted wake-sleep objective.

Algorithm 1 Variational Attention	**Algorithm 2** Variational Relaxed Attention
$\lambda \leftarrow \mathrm{enc}(x, \tilde{x}, y; \phi)$ ▷ *Compute var. params*	$\max_\theta \mathbb{E}_{z \sim p}[\log p(y \mid x, z)]$ ▷ *Pretrain fixed θ*
$z \sim q(z; \lambda)$ ▷ *Sample var. attention*	. . .
$\log f(x, z)$ ▷ Compute output dist	$u \sim \mathcal{U}$ ▷ *Sample unparam.*
$z' \leftarrow \mathbb{E}_{p(z' \mid x, \tilde{x})}[z']$ ▷ Compute soft atten.	$z \leftarrow g_\phi(u)$ ▷ *Reparam sample*
$B = \log f(x, z')$ ▷ Compute baseline dist	$\log f(x, z)$ ▷ Compute output dist
Backprop ∇_θ and ∇_ϕ based on eq. 1 and KL	Backprop ∇_θ and ∇_ϕ, reparam and KL

inference methods directly aim to tighten this gap. In particular, the ==*evidence lower bound* (ELBO)== is a parameterized bound over a family of distributions $q(z) \in \mathcal{Q}$ (with the constraint that the $\mathrm{supp}\, q(z) \subseteq \mathrm{supp}\, p(z \mid x, \tilde{x}, y)$),

$$\log \mathbb{E}_{z \sim p(z \mid x, \tilde{x})}[p(y \mid x, z)] \geq \mathbb{E}_{z \sim q(z)}[\log p(y \mid x, z)] - \mathrm{KL}[q(z) \,\|\, p(z \mid x, \tilde{x})]$$

This allows us to search over variational distributions q to improve the bound. It is tight when the variational distribution is equal to the posterior, i.e. $q(z) = p(z \mid x, \tilde{x}, y)$. Hard attention is a special case of the ELBO with $q(z) = p(z \mid x, \tilde{x})$.

There are many ways to optimize the evidence lower bound; an effective choice for deep learning applications is to use *amortized variational inference*. AVI uses an *inference network* to produce the parameters of the variational distribution $q(z; \lambda)$. The inference network takes in the input, query, and the output, i.e. $\lambda = enc(x, \tilde{x}, y; \phi)$. The objective aims to reduce the gap with the inference network ϕ while also training the generative model θ,

$$\max_{\phi, \theta} \mathbb{E}_{z \sim q(z; \lambda)}[\log p(y \mid x, z)] - \mathrm{KL}[q(z; \lambda) \,\|\, p(z \mid x, \tilde{x})]$$

With the right choice of optimization strategy and inference network this form of variational attention can provide a general method for learning latent alignment models. In the rest of this section, we consider strategies for accurately and efficiently computing this objective; in the next section, we describe instantiations of enc for specific domains.

Algorithm 1: Categorical Alignments First consider the case where $\mathcal{D}$, the alignment distribution, and $\mathcal{Q}$, the variational family, are categorical distributions. Here the generative assumption is that y is generated from a single index of x. Under this setup, a low-variance estimator of $\nabla_\theta \mathrm{ELBO}$, is easily obtained through a single sample from $q(z)$. For $\nabla_\phi \mathrm{ELBO}$, the gradient with respect to the KL portion is easily computable, but there is an optimization issue with the gradient with respect to the first term $\mathbb{E}_{z \sim q(z)}[\log f(x, z))]$.

Many recent methods target this issue, including neural estimates of baselines [43, 44], Rao-Blackwellization [51], reparameterizable relaxations [27, 41], and a mix of various techniques [63, 22]. We found that an approach using REINFORCE [66] along with a specialized baseline was effective. Formally, we first apply the likelihood-ratio trick to obtain an expression for the gradient with respect to the inference network parameters ϕ,

$$\nabla_\phi \mathbb{E}_{z \sim q(z)}[\log p(y \mid x, z)] = \mathbb{E}_{z \sim q(z)}[(\log f(x, z) - B)\nabla_\phi \log q(z)]$$

As with hard attention, we take a single Monte Carlo sample (now drawn from the variational distribution). Variance reduction of this estimate falls to the baseline term B. The ideal (and intuitive) baseline would be $\mathbb{E}_{z \sim q(z)}[\log f(x, z)]$, analogous to the value function in reinforcement learning. While this term cannot be easily computed, there is a natural, cheap approximation: soft attention (i.e. $\log f(x, \mathbb{E}[z])$). Then the gradient is

$$\mathbb{E}_{z \sim q(z)} \left[\left(\log \frac{f(x, z)}{f(x, \mathbb{E}_{z' \sim p(z' \mid x, \tilde{x})}[z'])} \right) \nabla_\phi \log q(z \mid x, \tilde{x}) \right] \tag{1}$$

Effectively this weights gradients to q based on the ratio of the inference network alignment approach to a soft attention baseline. Notably the expectation in the soft attention is over p (and not over q), and therefore the baseline is constant with respect to ϕ. Note that a similar baseline can also be used for hard attention, and we apply it to both variational/hard attention models in our experiments.

Algorithm 2: Relaxed Alignments Next consider treating both $\mathcal{D}$ and $\mathcal{Q}$ as Dirichlets, where z represents a mixture of indices. This model is in some sense closer to the soft attention formulation which assigns mass to multiple indices, though fundamentally different in that we still formally treat alignment as a latent variable. Again the aim is to find a low variance gradient for the expectation term of the objective under the variational distribution. Instead of using REINFORCE, certain continuous distributions allow the use reparameterization [31], where sampling $z \sim q(z)$ can be done by first sampling from a simple unparameterized distribution $\mathcal{U}$, and then applying a transformation $g_\phi(\cdot)$, yielding an unbiased estimator,

$$\mathbb{E}_{u \sim \mathcal{U}} \left[\nabla_\phi \log p(y|x, g_\phi(u)) \right] - \nabla_\phi \operatorname{KL} \left[q(z) \,\|\, p(z \,|\, x, \tilde{x}) \right]$$

The Dirichlet distribution is not directly reparameterizable. While transforming the standard uniform distribution with the inverse CDF of Dirichlet would result in a Dirichlet distribution, the inverse CDF does not have an analytical solution. However, we can use rejection based sampling to get a sample, and employ implicit differentiation to estimate the gradient by approximating the gradient of the CDF [28].

Empirically, we found the random initialization would result in convergence to uniform Dirichlet parameters for λ. (We suspect that it is easier to find low KL local optima towards the center of the simplex). In experiments, we therefore intialize the latent alignment model by first minimizing the Jensen bound, $\mathbb{E}_{z \sim p(z \,|\, x, \tilde{x})}[\log p(y \,|\, x, z)]$, with the same reparameterization methods, and then introducing the inference network.

4 Models and Methods

We experiment with variational attention in two different domains where attention-based models are essential and widely-used: neural machine translation and visual question answering.

Neural Machine Translation Neural machine translation (NMT) takes in a source sentence $w_1, \ldots w_T$ and predicts each word of a target sentence y_j in an auto-regressive manner. The model first contextually embeds each source word using a bidirectional LSTM to produce the vectors $x_1 \ldots x_T$. The query $\tilde{x}$ consists of an LSTM-based representation of the previous target words $y_{1:j-1}$. Attention is used to identify which source positions should be used to predict the target. The parameters of $\mathcal{D}$ are generated from an MLP between the query and source [6], and f concatenates the selected x_i with the query $\tilde{x}$ and passes it to an MLP to produce the distribution over the next target word y_j.

For variational attention, the inference network enc applies a bidirectional LSTM over the source and the target to obtain the hidden states $x_1, \ldots, x_T$ and $h_1, \ldots, h_S$, and produces the alignment scores at the j-th time step via a bilinear map, $s_i^{(j)} = \exp(h_j^\top \mathbf{U} x_i)$. For the categorical case, the scores are normalized, $q(z_i^{(j)} = 1) \propto s_i^{(j)}$; in the relaxed case the parameters of the Dirichlet are $\alpha_i^{(j)} = s_i^{(j)}$. Note, the inference network sees the entire target (through bidirectional LSTMs). The word embeddings are shared between the generative/inference networks, but the other parameters are separate.

Visual Question Answering Visual question answering (VQA) uses attention to locate the parts of an image that are necessary to answer a textual question. We follow the recently-proposed "bottom-up top-down" attention approach [2], which uses Faster R-CNN [52] to obtain object bounding boxes and performs mean-pooling over the convolutional features (from a pretrained ResNet-101 [25]) in each bounding box to obtain object representations $x_1, \ldots, x_T$. The query $\tilde{x}$ is obtained by running an LSTM over the question, the attention function a passes the query and the object representation through an MLP. The prediction function f is also similar to the NMT case: we concatenate the chosen x_i with the query $\tilde{x}$ to use as input to an MLP which produces a distribution over the output. The inference network enc uses the answer embedding h_y and combines it with x_i and $\tilde{x}$ to produce the variational (categorical) distribution,

$$q(z_i = 1) \propto \exp(u^\top \tanh(\mathbf{U}_1(x_i \odot \operatorname{ReLU}(\mathbf{V}_1 h_y)) + \mathbf{U}_2(\tilde{x} \odot \operatorname{ReLU}(\mathbf{V}_2 h_y))))$$

where $\odot$ is the element-wise product. This parameterization worked better than alternatives. We did not experiment with the relaxed case in VQA, as the object bounding boxes already give us the ability to attend to larger portions of the image.

Predictive Inference At test time, we need to marginalize out the latent variables, i.e. $\mathbb{E}_z[p(y \mid x, \tilde{x}, z)]$ using $p(z \mid x, \tilde{x})$, without access to q (as we do not know future words). In the categorical case, if speed is not an issue then simply enumerating alignments is preferable, which incurs a multiplicative cost of $O(T)$ (but the enumeration is parallelizable). Alternatively we experimented with a K-max renormalization, where we only take the top-K attention scores to obtain the attention distribution (by re-normalizing) to perform marginalization. This makes the multiplicative cost constant with respect to T. (We experiment with different values of K in the next section). For the relaxed case, sampling is necessary.

5 Experiments

Setup For NMT we use the IWSLT data set [11]. This dataset is relatively small, but has become a standard benchmark for experimental NMT models. We follow the same preprocessing as in [19] with the same Byte Pair Encoding vocabulary of 14,000 tokens [56]. For VQA, we use the VQA 2.0 dataset. As we are interested in intrinsic evaluation (i.e. log-likelihood) in addition to the standard VQA metric, we randomly select half of the standard validation set as the test set (since we need access to the actual labels).[7] (Therefore the numbers provided are not strictly comparable to existing work.) While the preprocessing is the same as [2], our numbers are worse than previously reported as we do not apply any of the commonly-utilized techniques to improve performance on VQA (e.g. augmenting the dataset with Visual Genome, using binary cross-entropy on soft labels instead of multiclass cross-entropy on hard labels, etc.).

Experiments vary three components of the systems: (a) training objective and model, (b) training approximations (comparing enumeration or sampling),[8] (c) test inference. All neural models have the same architecture and the exact same number of parameters θ (the inference network parameters ϕ vary, but are not used at test). When training hard and variational attention with sampling both use the same baseline, i.e the output from soft attention. The full architectures/hyperparameters for both NMT and VQA are given in Appendix B.

Results and Discussion Table 1 shows the main results for both NMT and VQA. We first note that hard attention underperforms soft attention, even when its expecation is enumerated. This indicates that Jensen's inequality alone is a poor bound. On the other hand, on both experiments, exact marginal likelihood performs better than soft attention, indicating that when possible to compute it is better to have latent alignments.

For NMT, variational attention with enumeration and sampling performs comparably to optimizing the explicit log marginal likelihood, despite the fact that it is optimizing a lower bound. We believe that this is due to the use of $q(z)$, which conditions on the entire source/target and therefore potentially provides better training signal to $p(z \mid x, \tilde{x})$ through the KL term. Note that it is also possible to have $q(z)$ come from a pretrained external model, such as a traditional alignment model [18]. Table 3 (left) shows these results in context compared to the best reported values for this task. Even with sampling, our system improves on the state-of-the-art. For VQA the trend is largely similar, and results for NLL with variational attention improve on soft attention and hard attention. However the task-specific evaluation metrics are slightly worse.

Table 2 (left) considers test inference for variational attention, comparing enumeration to K-max with $K = 5$. For all methods exact enumeration is better, however K-max is a reasonable approximation. Table 2 (right) shows the PPL of different models as we increase K. Note good performance requires $K > 1$, but that we observe only marginal benefits for $K > 5$. Finally, we observe that it is possible to *train* with soft attention and *test* using K-Max with a small performance drop (`Soft KMax` in Table 2 (right)). This possibly indicates that soft attention models are indeed approximating latent alignment models. (On the other hand, training with latent alignments and testing with soft attention performed badly).

Table 3 (right) looks at the entropy of the prior distribution learned by the different models. We see a significant range in the certainty of predictions. Notably hard attention has very low entropy

---

[7] VQA eval metric is defined as $\min\{\frac{\#\text{ humans that said answer}}{3}, 1\}$. Also note that since there are sometimes multiple answers for a given question, in such cases we sample (where the sampling probability is proportional to the number of humans that said the answer) to get a single label.

[8] Note that enumeration does not imply exact if we are enumerating an expectation on a lower bound.

Model	Objective	$\mathbb{E}$	NMT PPL	NMT BLEU	VQA NLL	VQA Eval
Soft Attention	$\log p(y \mid \mathbb{E}[z])$	-	7.03	32.31	1.76	58.93
Marginal Likelihood	$\log \mathbb{E}[p]$	Enum	6.33	33.08	1.69	60.33
Hard Attention	$\mathbb{E}_p[\log p]$	Enum	7.37	31.40	1.78	57.60
Hard Attention	$\mathbb{E}_p[\log p]$	Sample	7.38	31.00	1.82	56.30
Variational Relaxed Attention	$\mathbb{E}_q[\log p] - \text{KL}$	Sample	7.58	30.05	-	-
Variational Attention	$\mathbb{E}_q[\log p] - \text{KL}$	Enum	6.03	33.10	1.69	58.44
Variational Attention	$\mathbb{E}_q[\log p] - \text{KL}$	Sample	6.13	33.09	1.75	57.52

Table 1: Evaluation on neural machine translation (NMT) and visual question answering (VQA) for the various models. $\mathbb{E}$ column indicates whether the expectation is calculated via enumeration (Enum) or a single sample (Sample) during training. For NMT we evaluate intrinsically on perplexity (PPL) (lower is better) and extrinsically on BLEU (higher is better), where for BLEU we perform beam search with beam size 10 and length penalty (see Appendix B for further details). For VQA we evaluate intrinsically on negative log-likelihood (NLL) (lower is better) and extrinsically on VQA evaluation metric (higher is better). All results except for relaxed attention use enumeration at test time.

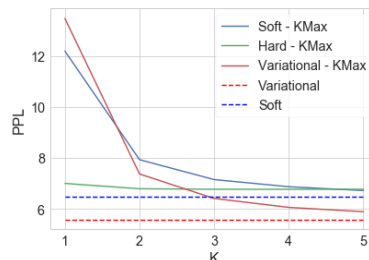| Model	PPL Exact	PPL K-Max	BLEU Exact	BLEU K-Max
Marginal Likelihood	6.33	6.89	33.08	32.97
Hard + Enum	7.37	7.37	31.40	31.37
Hard + Sample	7.38	7.38	31.00	31.04
Variational + Enum	6.03	6.37	33.10	33.00
Variational + Sample	6.13	6.45	33.09	33.01

Table 2: (Left) Performance change on NMT from exact decoding to K-Max decoding with $K = 5$. (see section 5 for definition of K-max decoding). (Right) Test perplexity of different approaches while varying the number of k-max samples to estimate $\mathbb{E}_z[p(y|x,\tilde{x})]$. Dotted lines compare soft baseline and variational with full enumeration.

(high certainty) whereas soft attention is quite high. The variational attention model falls in between. Figure 3 (left) illustrates the difference in practice.

Despite extensive experiments, we found that variational relaxed attention performed worse than other methods. In particular we found that when training with a Dirichlet KL, it is hard to reach low-entropy regions of the simplex, and so both models are more uniform than either soft or variational categorical attention. Table 3 (right) quantifies this issue. We experimented with other distributions such as Logistic-Normal and Gumbel-Softmax [27, 41] but neither fixed this issue. Others have also noted difficulty in training Dirichlet models with amortized inference [59].

Besides performance, an advantage of these models is the ability to perform posterior inference, since the q function can be used directly to obtain posterior alignments. Contrast this with hard attention where $q = p(z \mid x, \tilde{x})$, i.e. the variational posterior is independent of the future information. Figure 3 shows the alignments of p and q for variational attention over a fixed sentence (see Appendix C for more examples). We see that q is able to use future information to correct alignments. We note that the inability of soft and hard attention to produce good alignments has been noted as a major issue in NMT [33]. While q is not used directly in left-to-right NMT decoding, it could be employed for other applications such as in an iterative refinement approach [48, 36].

Potential Limitations While this technique is a promising alternative to soft attention, there are some practical limitations: (a) Variational/hard attention needs a good baseline estimator in the form of soft attention. We found this to be a necessary component for adequately training the system. This may prevent this technique from working when T is intractably large and soft attention is not an option. (b) For some applications, the model relies heavily on having a good posterior estimator. In VQA we had to utilize domain structure for the enc function. (c) Recent models such as the Transformer [64], utilize many repeated attention models. For instance the current best translation
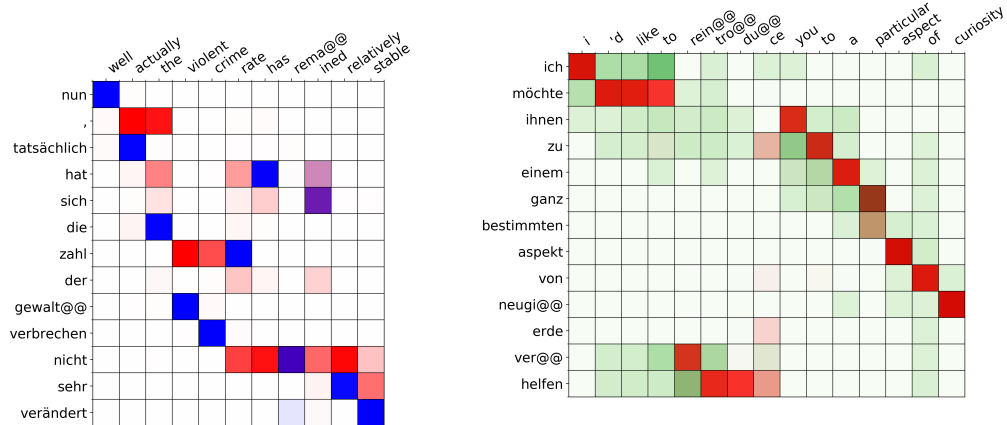
Figure 3: (Left) An example demonstrating the difference between the prior alignment (red) and the variational posterior (blue) when translating from DE-EN (left-to-right). Note the improved blue alignments for `actually` and `violent` which benefit from seeing the next word. (Right) Comparison of soft attention (green) with the p of variational attention (red). Both models imply a similar alignment, but variational attention is lower entropy.

Model	BLEU
Beam Search Optimization [67]	26.36
Actor-Critic [5]	28.53
Neural PBMT + LM [26]	30.08
Minimum Risk Training [19]	32.84
Soft Attention	32.31
Marginal Likelihood	33.08
Hard Attention + Enum	31.40
Hard Attention + Sample	30.42
Variational Relaxed Attention	30.05
Variational Attention + Enum	33.10
Variational Attention + Sample	33.09

Model	Entropy	
	NMT	VQA
Soft Attention	1.24	2.70
Marginal Likelihood	0.82	2.66
Hard Attention + Enum	0.05	0.73
Hard Attention + Sample	0.07	0.58
Variational Relaxed Attention	2.02	-
Variational Attention + Enum	0.54	2.07
Variational Attention + Sample	0.52	2.44

Table 3: (Left) Comparison against the best prior work for NMT on the IWSLT 2014 German-English test set. (Right) Comparison of different models in terms of implied discrete entropy (lower = more peaked alignment).

models have the equivalent of 150 different attention queries per word translated. It is unclear if this approach can be used at that scale.

6 Conclusion

Attention methods are ubiquitous tool for areas like natural language processing; however they are difficult to use as latent variable models. This work explores alternative approaches to latent alignment, through variational attention with promising result. Future work will experiment with scaling the method on larger-scale tasks and in more complex models, such as multi-hop attention models, transformer models, and structured models, as well as utilizing these latent variables for interpretability and as a way to incorporate prior knowledge.

Acknowledgements

References

[1] David Alvarez-Melis and Tommi S Jaakkola. A Causal Framework for Explaining the Predictions of Black-Box Sequence-to-Sequence Models. In *Proceddings of EMNLP*, 2017.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of CVPR*, 2018.

[3] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple Object Recognition with Visual Attention. In *Proceedings of ICLR*, 2015.

[4] Jimmy Ba, Ruslan R Salakhutdinov, Roger B Grosse, and Brendan J Frey. Learning Wake-Sleep Recurrent Attention Models. In *Proceedings of NIPS*, 2015.

[5] Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. An Actor-Critic Algorithm for Sequence Prediction. In *Proceedings of ICLR*, 2017.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of ICLR*, 2015.

[7] Hareesh Bahuleyan, Lili Mou, Olga Vechtomova, and Pascal Poupart. Variational Attention for Sequence-to-Sequence Models. *arXiv:1712.08207*, 2017.

[8] Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational linguistics*, 19(2):263–311, 1993.

[9] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguist.*, 19(2):263–311, June 1993.

[10] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance Weighted Autoencoders. In *Proceedings of ICLR*, 2015.

[11] Mauro Cettolo, Jan Niehues, Sebastian Stuker, Luisa Bentivogli, and Marcello Federico. Report on the 11th IWSLT evaluation campaign. In *Proceedings of IWSLT*, 2014.

[12] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, Attend and Spell. *arXiv:1508.01211*, 2015.

[13] Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. Describing Multimedia Content using Attention-based Encoder-Decoder Networks. In *IEEE Transactions on Multimedia*, 2015.

[14] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-Based Models for Speech Recognition. In *Proceedings of NIPS*, 2015.

[15] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron Courville, and Yoshua Bengio. A Recurrent Latent Variable Model for Sequential Data. In *Proceedings of NIPS*, 2015.

[16] Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. Incorporating Structural Alignment Biases into an Attentional Neural Translation Model. In *Proceedings of NAACL*, 2016.

[17] Yuntian Deng, Anssi Kanervisto, Jeffrey Ling, and Alexander M Rush. Image-to-Markup Generation with Coarse-to-Fine Attention. In *Proceedings of ICML*, 2017.

[18] Chris Dyer, Victor Chahuneau, and Noah A. Smith. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of NAACL*, 2013.

[19] Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. Classical Structured Prediction Losses for Sequence to Sequence Learning. In *Proceedings of NAACL*, 2018.

[20] Marco Fraccaro, Soren Kaae Sonderby, Ulrich Paquet, and Ole Winther. Sequential Neural Models with Stochastic Layers. In *Proceedings of NIPS*, 2016.

[21] Anirudh Goyal, Alessandro Sordoni, Marc-Alexandre Cote, Nan Rosemary Ke, and Yoshua Bengio. Z-Forcing: Training Stochastic Recurrent Networks. In *Proceedings of NIPS*, 2017.

[22] Will Grathwohl, Dami Choi, Yuhuai Wu, Geoffrey Roeder, and David Duvenaud. Backpropagation through the Void: Optimizing control variates for black-box gradient estimation. In *Proceedings of ICLR*, 2018.

[23] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. Incorporating Copying Mechanism in Sequence-to-Sequence Learning. 2016.

[24] Caglar Gulcehre, Sarath Chandar, Kyunghyun Cho, and Yoshua Bengio. Dynamic Neural Turing Machine with Soft and Hard Addressing Schemes. *arXiv:1607.00036*, 2016.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of CVPR*, 2016.

[26] Po-Sen Huang, Chong Wang, Sitao Huang, Dengyong Zhou, and Li Deng. Towards neural phrase-based machine translation. In *Proceedings of ICLR*, 2018.

[27] Eric Jang, Shixiang Gu, and Ben Poole. Categorical Reparameterization with Gumbel-Softmax. In *Proceedings of ICLR*, 2017.

[28] Martin Jankowiak and Fritz Obermeyer. Pathwise Derivatives Beyond the Reparameterization Trick. In *Proceedings of ICML*, 2018.

[29] Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. Structured Attention Networks. In *Proceedings of ICLR*, 2017.

[30] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of ICLR*, 2015.

[31] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *Proceedings of ICLR*, 2014.

[32] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics, 2007.

[33] Philipp Koehn and Rebecca Knowles. Six Challenges for Neural Machine Translation. *arXiv:1706.03872*, 2017.

[34] Rahul G. Krishnan, Uri Shalit, and David Sontag. Structured Inference Networks for Nonlinear State Space Models. In *Proceedings of AAAI*, 2017.

[35] Dieterich Lawson, Chung-Cheng Chiu, George Tucker, Colin Raffel, Kevin Swersky, and Navdeep Jaitly. Learning Hard Alignments in Variational Inference. In *Proceedings of ICASSP*, 2018.

[36] Jason Lee, Elman Mansimov, and Kyunghyun Cho. Deterministic Non-Autoregressive Neural Sequence Modeling by Iterative Refinement. *arXiv:1802.06901*, 2018.

[37] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing Neural Rredictions. In *Proceedings of EMNLP*, 2016.

[38] Yang Liu and Mirella Lapata. Learning Structured Text Representations. In *Proceedings of TACL*, 2017.

[39] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of EMNLP*, 2015.

[40] Xuezhe Ma, Yingkai Gao, Zhiting Hu, Yaoliang Yu, Yuntian Deng, and Eduard Hovy. Dropout with Expectation-linear Regularization. In *Proceedings of ICLR*, 2017.

[41] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *Proceedings of ICLR*, 2017.

[42] André F. T. Martins and Ramón Fernandez Astudillo. From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification. In *Proceedings of ICML*, 2016.

[43] Andriy Mnih and Karol Gregor. Neural Variational Inference and Learning in Belief Networks. In *Proceedings of ICML*, 2014.

[44] Andriy Mnih and Danilo J. Rezende. Variational Inference for Monte Carlo Objectives. In *Proceedings of ICML*, 2016.

[45] Volodymyr Mnih, Nicola Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent Models of Visual Attention. In *Proceedings of NIPS*, 2015.

[46] Vlad Niculae and Mathieu Blondel. A Regularized Framework for Sparse and Structured Neural Attention. In *Proceedings of NIPS*, 2017.

[47] Vlad Niculae, André F. T. Martins, Mathieu Blondel, and Claire Cardie. SparseMAP: Differentiable Sparse Structured Inference. In *Proceedings of ICML*, 2018.

[48] Roman Novak, Michael Auli, and David Grangier. Iterative Refinement for Machine Translation. *arXiv:1610.06602*, 2016.

[49] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of EMNLP*, 2014.

[50] Colin Raffel, Minh-Thang Luong, Peter J Liu, Ron J Weiss, and Douglas Eck. Online and Linear-Time Attention by Enforcing Monotonic Alignments. In *Proceedings of ICML*, 2017.

[51] Rajesh Ranganath, Sean Gerrish, and David M. Blei. Black Box Variational Inference. In *Proceedings of AISTATS*, 2014.

[52] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of NIPS*, 2015.

[53] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proceedings of ICML*, 2014.

[54] Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomas Kocisky, and Phil Blunsom. Reasoning about Entailment with Neural Attention. In *Proceedings of ICLR*, 2016.

[55] Alexander M. Rush, Sumit Chopra, and Jason Weston. A Neural Attention Model for Abstractive Sentence Summarization. In *Proceedings of EMNLP*, 2015.

[56] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of ACL*, 2016.

[57] Iulian Vlad Serban, Alessandro Sordoni, Laurent Charlin Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *Proceedings of AAAI*, 2017.

[58] Bonggun Shin, Falgun H Chokshi, Timothy Lee, and Jinho D Choi. Classification of Radiology Reports Using Neural Attention Models. In *Proceedings of IJCNN*, 2017.

[59] Akash Srivastava and Charles Sutton. Autoencoding Variational Inference for Topic Models. In *Proceedings of ICLR*, 2017.

[60] Jinsong Su, Shan Wu, Deyi Xiong, Yaojie Lu, Xianpei Han, and Biao Zhang. Variational Recurrent Neural Machine Translation. In *Proceedings of AAAI*, 2018.

[61] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-To-End Memory Networks. In *Proceedings of NIPS*, 2015.

[62] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. Modeling Coverage for Neural Machine Translation. In *Proceedings of ACL*, 2016.

[63] George Tucker, Andriy Mnih, Chris J. Maddison, Dieterich Lawson, and Jascha Sohl-Dickstein. REBAR: Low-variance, Unbiased Gradient Estimates for Discrete Latent Variable Models. In *Proceedings of NIPS*, 2017.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *Proceedings of NIPS*, 2017.

[65] Stephan Vogel, Hermann Ney, and Christoph Tillmann. HMM-based Word Alignment in Statistical Translation. In *Proceedings of COLING*, 1996.

[66] Ronald J. Williams. Simple Statistical Gradient-following Algorithms for Connectionist Reinforcement Learning. *Machine Learning*, 8, 1992.

[67] Sam Wiseman and Alexander M. Rush. Sequence-to-Sequence learning as Beam Search Optimization. In *Proceedings of EMNLP*, 2016.

[68] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Klaus Macherey Qin Gao, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, Nishant Patil George Kurian, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv:1609.08144*, 2016.

[69] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of ICML*, 2015.

[70] Zichao Yang, Kiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked Attention Networks for Image Question Answering. In *Proceedings of CVPR*, 2016.

[71] Lei Yu, Phil Blunsom, Chris Dyer, Edward Grefenstette, and Tomas Kocisky. The Neural Noisy Channel. In *Proceedings of ICLR*, 2017.

[72] Lei Yu, Jan Buys, and Phil Blunsom. Online Segment to Segment Neural Transduction. In *Proceedings of EMNLP*, 2016.

[73] Biao Zhang, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. Variational Neural Machine Translation. In *Proceedings of EMNLP*, 2016.

[74] Chen Zhu, Yanpeng Zhao, Shuaiyi Huang, Kewei Tu, and Yi Ma. Structured Attentions for Visual Question Answering. In *Proceedings of ICCV*, 2017.

Supplementary Materials for
Latent Alignment and Variational Attention

Appendix A: Proof of Proposition 1

Proposition. *Define $g_{x,\hat{y}} : \Delta^{T-1} \mapsto [0,1]$ to be the function given by $g_{x,\hat{y}}(z) = f(x,z)_{\hat{y}}$ (i.e. $g_{x,\hat{y}}(z) = p(y = \hat{y} \mid x, \tilde{x}, z))$ for a twice differentiable function f. Let $H_{g_{x,\hat{y}}}(z)$ be the Hessian of $g_{x,\hat{y}}(z)$ evaluated at z, and further suppose $\|H_{g_{x,\hat{y}}}(z)\|_2 \leq c$ for all $z \in \Delta^{T-1}, \hat{y} \in \mathcal{Y}\}$, and x, where $\|\cdot\|_2$ is the spectral norm. Then for all $\hat{y} \in \mathcal{Y}$,*

$$|\, p(y = \hat{y} \mid x, \tilde{x}) - p_{\text{soft}}(y = \hat{y} \mid x, \tilde{x})\,| \leq c$$

Proof. We begin by performing Taylor's expansion of $g_{x,\hat{y}}$ at $\mathbb{E}[z]$:

$$\mathbb{E}[g_{x,\hat{y}}(z)] = \mathbb{E}\Big[g_{x,\hat{y}}(\mathbb{E}[z]) + (z - \mathbb{E}[z])^\top \nabla g_{x,\hat{y}}(\mathbb{E}[z]) + \frac{1}{2}(z - \mathbb{E}[z])^\top H_{g_{x,\hat{y}}}(\hat{z})(z - \mathbb{E}[z])\Big]$$

$$= g_{x,\hat{y}}(\mathbb{E}[z]) + \frac{1}{2}\mathbb{E}[(z - \mathbb{E}[z])^\top H_{g_{x,\hat{y}}}(\hat{z})(z - \mathbb{E}[z])]$$

for some $\hat{z} = \lambda z + (1 - \lambda)\mathbb{E}[z], \lambda \in [0,1]$. Then letting $u = z - \mathbb{E}[z]$, we have

$$|\,(z - \mathbb{E}[z])^\top H_{g_{x,\hat{y}}}(\hat{z})(z - \mathbb{E}[z])\,| = |\,\|u\|_2^2 \frac{u^\top}{\|u\|_2} H_{g_{x,\hat{y}}}(\hat{z}) \frac{u}{\|u\|_2}\,|$$

$$\leq \|u\|_2^2\, c$$

where $c = \max\{|\lambda_{\max}|, |\lambda_{\min}|\}$ is the largest absolute eigenvalue of $H_{g_{x,\hat{y}}}(\hat{z})$. (Here $\lambda_{\max}$ and $\lambda_{\min}$ are maximum/minimum eigenvalues of $H_{g_{X,q}}(\hat{z})$). Note that c is also equal to the spectral norm $\|H_{g_{X,q}}(\hat{z})\|_2$ since the Hessian is symmetric.

Then,

$$|\,\mathbb{E}[(z - \mathbb{E}[z])^\top H_{g_{x,\hat{y}}}(\hat{z})(z - \mathbb{E}[z])]\,| \leq \mathbb{E}[\,|(z - \mathbb{E}[z])^\top H_{g_{x,\hat{y}}}(\hat{z})(z - \mathbb{E}[z])\,|]$$

$$\leq \mathbb{E}[\|u\|_2^2 c]$$

$$\leq 2c$$

Here the first inequality follows due to the convexity of the absolute value function and the last inequality follows since

$$\|u\|_2^2 = (z - \mathbb{E}[z])^\top (z - \mathbb{E}[z])$$

$$= z^\top z + \mathbb{E}[z]^\top \mathbb{E}[z] - 2\mathbb{E}[z]^\top z$$

$$\leq z^\top z + \mathbb{E}[z]^\top \mathbb{E}[z]$$

$$\leq 2$$

where the last two inequalities are due to the fact that $z, \mathbb{E}[z] \in \Delta^{T-1}$. Then putting it all together we have,

$$|\, p(y = \hat{y} \mid x, \tilde{x}) - p_{\text{soft}}(y = \hat{y} \mid x, \tilde{x})\,| = |\,\mathbb{E}[g_{x,\hat{y}}(z)] - g_{x,\hat{y}}(\mathbb{E}[z])\,|$$

$$= \frac{1}{2}|\,\mathbb{E}[(z - \mathbb{E}[z])^\top H_{g_{x,\hat{y}}}(\hat{z})(z - \mathbb{E}[z])]\,|$$

$$\leq c$$

$\square$

14

Appendix B: Experimental Setup

Neural Machine Translation

For data processing we closely follow the setup in [19], which uses Byte Pair Encoding over the combined source/target training set to obtain a vocabulary size of 14,000 tokens. However, different from [19] which uses maximum sequence length of 175, for faster training we only train on sequences of length up to 125.

The encoder is a two-layer bi-directional LSTM with 512 units in each direction, and the decoder as a two-layer LSTM with with 768 units. For the decoder, the convex combination of source hidden states at each time step from the attention distribution is used as additional input at the next time step. Word embedding is 512-dimensional.

The inference network consists of two bi-directional LSTMs (also two-layer and 512-dimensional each) which is run over the source/target to obtain the hidden states at each time step. These hidden states are combined using bilinear attention [39] to produce the variational parameters. (In contrast the generative model uses MLP attention from [6], though we saw little difference between the two parameterizations). Only the word embedding is shared between the inference network and the generative model.

Other training details include: batch size of 6, dropout rate of 0.3, parameter initialization over a uniform distribution $\mathcal{U}[-0.1, 0.1]$, gradient norm clipping at 5, and training for 30 epochs with Adam (learning rate = 0.0003, $\beta_1 = 0.9$, $\beta_2 = 0.999$) [30] with a learning rate decay schedule which starts halving the learning rate if validation perplexity does not improve. Most models converged well before 30 epochs.

For decoding we use beam search with beam size 10 and length penalty $\alpha = 1$, from [68]. The length penalty added about 0.5 BLEU points across all the models.

Visual Question Answering

The model first obtains object features by mean-pooling the pretrained ResNet-101 features [25] (which are 2048-dimensional) over object regions given by Faster R-CNN [52].The ResNet features are kept fixed and not fine-tuned during training. We fix the maximum number of possible regions to be 36. For the question embedding we use a one-layer LSTM with 1024 units over word embeddings. The word embeddings are 300-dimensional and initialized with GloVe [49]. The generative model produces a distribution over the possible objects via applying MLP attention, i.e.

$$p(z_i = 1 \,|\, x, \tilde{x}) \propto \exp(w^\top \tanh(\mathbf{W}_1 x_i + \mathbf{W}_2 \tilde{x}))$$

The selected image region is concatenated with the question embedding and fed to a one-layer MLP with ReLU non-linearity and 1024 hidden units.

The inference network produces a categorical distribution over the image regions by interacting the answer embedding h_y (which are 256-dimensional and initialized randomly) with the question embedding $\tilde{x}$ and the image regions x_i,

$$q(z_i = 1) \propto \exp(u^\top \tanh(\mathbf{U}_1(x_i \odot \mathrm{ReLU}(\mathbf{V}_1 h_y)) + \mathbf{U}_2(\tilde{x} \odot \mathrm{ReLU}(\mathbf{V}_2 h_y))))$$

where $\odot$ denotes element-wise multiplication. The generative/inference attention MLPs have 1024 hidden units each (i.e. $w, u \in \mathbb{R}^{1024}$).

Other training details include: batch size of 512, dropout rate of 0.5 on the penultimate layer (i.e. before affine transformation into answer vocabulary), and training for 50 epochs with with Adam (learning rate = 0.0005, $\beta_1 = 0.9$, $\beta_2 = 0.999$) [30].

In cases where there is more than one answer for a given question/image pair, we randomly sample the answer, where the sampling probability is proportional to the number of humans who gave the answer.

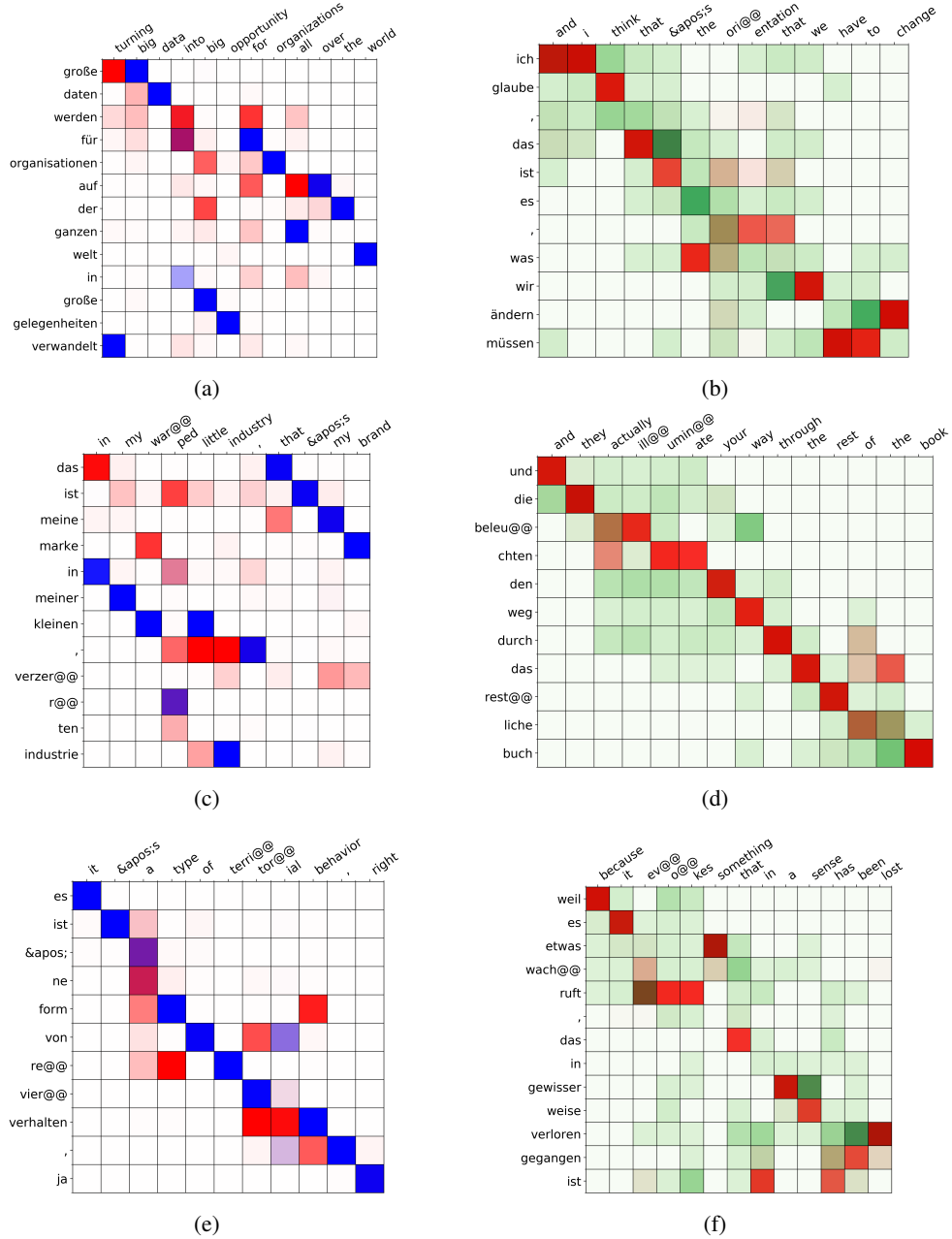# Appendix C: Additional Visualizations

(a)

(b)

(c)

(d)

(e)

(f)

Figure 4: (Left Column) Further examples highlighting the difference between the prior alignment (red) and the variational posterior (blue) when translating from DE-EN (left-to-right). The variational posterior is able to better handle reordering; in (a) the variational posterior successfully aligns 'turning' to 'verwandelt', in (c) we see a similar pattern with the alignment of the clause 'that's my brand' to 'das ist meine marke'. In (e) the prior and posterior both are confused by the '-ial' in 'territor-ial', however the posterior still remains more accurate overall and correctly aligns the rest of 'revierverhalten' to 'territorial behaviour'. (Right Column) Additional comparisons between soft attention (green) and the prior alignments of variational attention (red). Alignments from both models are similar, but variational attention is lower entropy. Both soft and variational attention rely on aligning the inserted English word 'orientation' to the comma in (b) since a direct translation does not appear in the German source.